

SPARK Conditional Probability

method map(row)

```
list <- []
```

```
movie_dict <- {}
```

```
movie_list <- split the row
```

```
sort movie_list
```

```
for all movie_i in movie_list do
```

```
    for all movie_j > movie_i in movie_list do
```

```
        if movie_j in movie_dict do
```

```
            movie_dict[movie_j] <- movie_dict[movie_j] + 1
```

```
        else
```

```
            movie_dict[movie_j] <- 1
```

```
    list.append((movie_i, movie_dict))
```

```
    movie_dict <- {}
```

```
return list
```

method reduce(movie_dict, new_movie_dict)

```
for all movie in new_movie_dict do
```

```
    if movie in movie_dict do
```

```
        movie_dict[movie] <- movie_dict[movie] + 1
```

```
    else
```

```
        movie_dict[movie] <- 1
```

```
return movie_dict
```

method filter(row)

```
movie_i, movie_dict <- data
```

```
list <- []
```

```

mcount <- 0
for all movie in movie_dict do
    mcount <- mcount + movie_dict[movie]
movie_name_1 <- global_movie_names[movie_i]
for all movie in movie_dict do
    conditional_probability <- movie_dict[movie]/mcount
    if conditional_probability > threshold:
        movie_name_2 <- global_movie_names[movie]
        list.append((movie_name_1, movie_name_2,
                    conditional_probability))

return list

```

method load_movie_names()

read movies.csv file and store the movie names into a global dictionary
 where movie_id is key and movie_name is the value

method main()

```

read text file as RDD
RDD.map(split row into user_id, movie_id, rating)
RDD.filter(rows with rating >= 4.0)
RDD.map(row <- (user_id, movie_id))
RDD.map(for each user_id concatenate all movie_id's)
RDD.map(row <- concatenated movie_id's)
RDD.map(map)
RDD.reduceByKey(merge the stripes for each movie)
global_movie_names <- load_movie_names()
RDD.map(filter)
save RDD as text file

```